Applying the Rasch model in the validation of a test of metalinguistic knowledge

Applying the Rasch Model in the Validation of a Test of Metalinguistic Knowledge

Brian Wistner

Abstract

The purpose of the current study was to examine the construct validity and unidimensionality of a test of first language metalinguistic knowledge. A ten-item multiple-choice test of Japanese metalinguistic terminology was designed and administered to 100 Japanese university students. The results of a Rasch analysis indicated that the calibrated item estimates were sufficiently accurate and precise—all of the items exhibited good fit with the Rasch model. Implications for test revision focused on the possibility of revising some items and increasing the number of items to provide fuller construct coverage and higher reliability and separation estimates.

Second language (L2) researchers have defined and tested metalinguistic knowledge in numerous ways. Previous studies, however, have often conceptualized and operationalized metalinguistic knowledge as knowledge of L2 metalinguistic terms that were tested in the L2. These operationalizations are problematic on a number of fronts. Primarily, by operationalizing metalinguistic knowledge through tests of L2 metalinguistic terms and rules, metalinguistic knowledge could be confounded with L2 proficiency. The purpose of the current study was to create a test of metalinguistic knowledge in the participants' first language (L1), and to collect evidence to support the construct validity and unidimensionality of the test through a Rasch analysis.

Previous Studies of Metalinguistic Knowledge

Previous studies have differed in their use of L1 or L2 metalinguistic tests or a combination of both. Of importance here is the plausible interaction among learners' L1, the language of the metalinguistic knowledge tests, and the content of the tests.

In an oft-cited study of metalinguistic knowledge, Alderson, Clapham, and Steel (1997) examined the relationship among metalinguistic knowledge, language aptitude, and language proficiency. Five-hundred-nine native English speaking university students completed a metalinguistic knowledge test that consisted of three parts: Section 1 asked learners to identify parts of speech in English and French; section 2 asked learners to judge the acceptability of French sentences and to state the rules which were broken; section 3 asked learners to make corrections to erroneous English sentences and to state the grammatical rules that had been violated. The mean scores on the metalinguistic tests were between 33% and 65%. The authors viewed this result as an indication that most learners had difficulty with

providing or identifying metalinguistic terms or with providing explanations of rules. Of interest to the current study was the result that learners were better able to correct errors in French than they were able to state rules that had been broken or to use metalinguistic terminology to explain the violated rules. While the researchers questioned the relative efficacy of metalanguage, another explanation for the results could be in the role of L2 proficiency of the participants and in test method effects.

In a study that focused on first-year undergraduate students' metalinguistic knowledge, Elder, Warren, Hajek, Manwaring, and Davies (1999) examined the relationship between levels of metalinguistic knowledge and language learning in university. Native English speakers studying a variety of L2s took a metalinguistic knowledge test in English, and L2 French learners also took a metalinguistic knowledge test in French (n = 87) and various French proficiency tests (n = 32).

Comparisons of the scores on the various tests revealed that first-year university students tended to have low levels of metalinguistic knowledge. Grammatical terms such as *subject*, *noun*, *verb*, and *adjective* were relatively easy for the sample, but other items such as *predicate* were difficult. Similar results were reported for the L2 French learners' levels of L2 French metalinguistic knowledge. Learners were better able to identify examples of metalinguistic terms than they were able to use the terms to explain grammatical rules. Statistically significant correlations were found for the advanced learners of L2 French on the French metalinguistic knowledge test (r = .56) and the English knowledge test (r = .43). The results were interpreted as indicating that as learners' proficiency levels increase, so do their levels of metalinguistic knowledge.

Ellis (2005) administered a metalinguistic knowledge test to a group of learners with mixed L2 proficiency (n = 91). This test included multiple-choice questions about metalinguistic terminology and questions related to the identification of target structures in a reading passage. The identification of structures in a reading passage based on a list of metalinguistic terminology has been used comparatively less frequently than tests consisting of multiple-choice questions and has the added benefit of a reduction in the effectiveness of guessing.

Other abilities have been included in some definitions of L2 metalinguistic knowledge. For instance, Roehr (2007) included a measure of L2 language-analytic ability in her definition of metalinguistic ability. A multiple-choice L2 German proficiency test (n = 32), a L2 metalinguistic knowledge test (n = 54), and a L2 language-analytic ability test (n = 54) were administered to learners of L2 German. Scores on the proficiency and metalinguistic knowledge tests were statistically significantly correlated, and a one-factor solution explaining 82 percent of the variance was found through a principle components analysis.

In summary, in some previous studies, metalinguistic knowledge has been operationalized through the use of tests given in the participants' L1 and L2, while many influential studies have tested metalinguistic knowledge in the participants' L2, which casts doubt on the interpretations drawn from those studies regarding the unidimensionality of proficiency and metalinguistic knowledge. To tease apart the latent structure of these complex and related constructs, a test of metalinguistic knowledge given in the participants' L1 is needed. To this end, a test of Japanese metalinguistic knowledge was created, administered, and the results were analyzed. The purpose of the current study was to investigate the extent to which the test items are useful for measuring Japanese university students'

levels of metalinguistic knowledge.

Research Questions

The following research question guided the current study: To what extent do the test items conform to the expectations of the Rasch model? More specifically, the present study sought to determine (a) the degree to which the test items exhibited construct coverage along the Rasch dimension; (b) the extent to which the test items fit the Rasch model; and (c) how precise the item difficulty estimates were.

Method

Participants

One-hundred Japanese university students participated in the study. Twenty-eight were second-year students, 58 were third-year students, and 14 were fourth-year students. All of the participants had attended Japanese junior and senior high schools at which they were exposed to English instruction that included a high usage of metalinguistic terminology in Japanese.

Materials

In addition to a background questionnaire, a ten-item Japanese test of metalinguistic knowledge was created to test Japanese university students' levels of L1 representations of metalinguistic knowledge about English. Experienced Japanese teachers of English ranked a list of Japanese metalinguistic terms and English rules in terms of perceived difficulty. Based on these expert opinions, rules and terms were chosen as candidates for inclusion on the test. Multiple choice test items were then created for ten of the target terms and rules. The majority of the items required participants to identify the correct metalinguistic term that described or explained the structures or words included in English sentences. All of the items included only metalinguistic terminology in Japanese. An English translation of an item targeting identification of interrogative adverbs was as follows:

1. *When does the show begin?*

   What grammatical terminology can be used to explain this *when*?

   a. interrogative adverb
   b. interrogative pronoun
   c. object of the verb
   d. interrogative adjective

Analysis

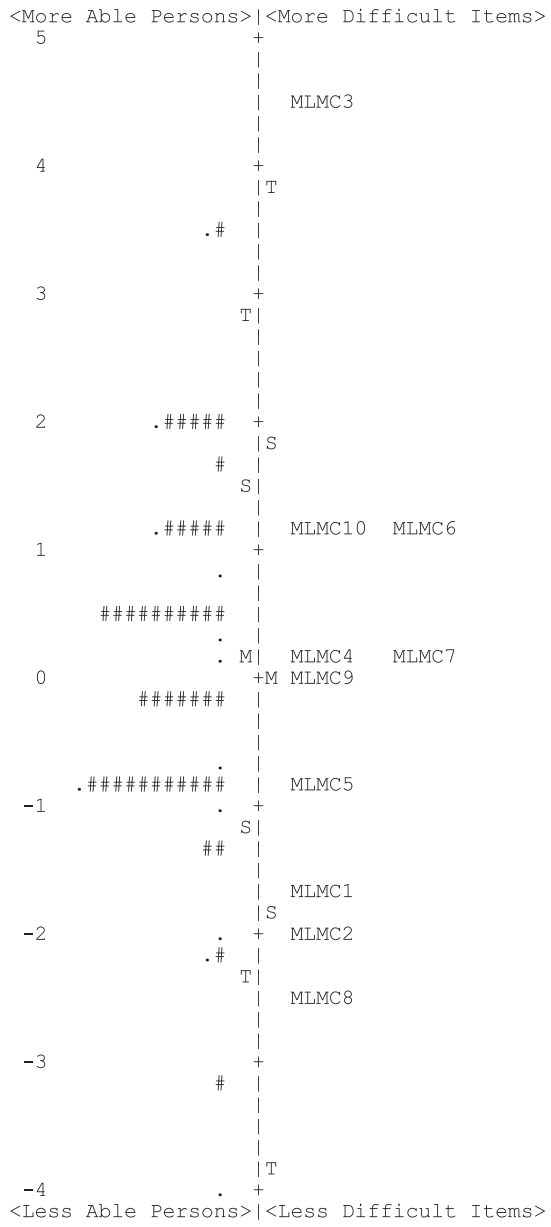
The ten metalinguistic knowledge test items were subjected to a Rasch analysis. First, descriptive statistics were examined to assess the general distribution of measures, model error, and the overall reliability of the person estimates. Second, the test items were calibrated to gain insight into the distribution of item difficulty estimates, item error estimates, and item fit statistics. All analyses were carried out in Winsteps 3.67.0.

Results and Discussion

The average of the person measures (M = .18; SD = 1.29) was slightly higher than the arbitrarily set mean of the item estimates (M = .00; SD = 1.91). This value implies that this set of items was relatively easy on average for this sample of participants; however, .18 is less than a quarter of a logit from the mean of the item difficulty estimates, which could be interpreted as fairly well-targeted. Both the person ability estimates and the item difficulty estimates spanned a wide range: The range of the person ability estimates was 6.68 logits (Max: 3.52; Min: -3.16), and the range of item difficulty estimates was 7.00 logits (Max: 4.43; Min: -2.57). A visual representation of the distribution of person ability estimates and item difficulty estimates is shown in Figure 1. Most notable when examining this figure are the gaps between many of the item difficulty estimates. Item 3 was estimated at 4.43 logits, but there were no test takers estimated to be in that range of ability. This item's location could be viewed as acceptable in that item 3 provided relevant information for the measurement of even the most able participants. More problematic, however, was the over three-logit gap between item 3 (4.43 logits) and item 6 (1.15 logits). Furthermore, approximately 0.5 to 1.00-logit gaps existed throughout the range of the other eight items. Moreover, items 4, 7, and 9 were estimated to be near or at the mean of the person ability estimates, but those items were tightly grouped, which could negatively affect the measurement and separation of the majority of the participants who were grouped around the mean. Ideally, those items would have been staggered along the continuum of item difficulty estimates, increasing from the lower ranges by .20 logits so that person abilities could be estimated with greater precision.

The exact item difficulty estimates are shown in Table 1. The precision of the item difficulty estimates varied depending on the location of the item. Items estimated to be at the higher and lower ends of the Rasch variable had higher error estimates due to the lack of information received about those item. Few participants were estimated to fall in the same ranges as those items, resulting in less measurement precision and measurement-relevant information. Six of the ten items had error estimates at or below .25 logits. Wright (1977) suggested that for tests of 20 or 30 items calibrated on a sample of 100 participants, error estimates around the mean of .25 or lower contribute positively to the measurement of the sample. As the current test had only 10 items, the error estimates for the items around the mean could be viewed as acceptable, and that those items are providing useful information for measurement.

Table 1 also shows the fit statistics for the items. Infit mean-square statistics should be centered on 1.00 for items that fit the Rasch model. Fit statistics over 1.00 (i.e., underfit) represent more variation in the responses to that item than predicted by the Rasch model, while figures under 1.00 represent overfit, which is less detrimental to measurement than underfit. Using the guideline of two standard deviations above or below the mean of infit mean-square statistics to judge goodness-of-fit to the Rasch model, none of the items exhibited misfit. Figure 2 shows a plot of the infit statistics for the ten items. While a few items slightly overfit the model, these items were not so concerning due to the relatively minor deviations from expected model fit. Items showing a high degree of overfit do not adversely affect the measurement of the sample—too much overfit indicates an intensification of the Rasch variable (i.e., the responses are overly deterministic). Item 10 displayed the most underfit of the items (infit MNSQ =



Note. Each # represents approximately two persons. MLMC stands for *metalinguistic multiple choice*, which is followed by the item number.

Figure 1. Item-map of the 10 metalinguistic test items.

1.16), but this figure was within two standard deviations of the mean and was not statistically significant (*Z-STD* = 1.4). Furthermore, the point-biserial correlation coefficients provided additional evidence of fit to the Rasch model. All of the coefficients were positive, and seven of the ten coefficients were over .40, indicating strong relationships between answers to each of those items and overall test scores. Thus, the ten items could be considered accurate, and there was statistical evidence that they were targeting

the latent variable.

The results of the Rasch analysis provide clear insights as to how this test could be revised. First, some of the items that were grouped together (i.e., items 6 and 10; items 4, 7, and 9) should be rewritten to adjust their difficulty estimates so that they are evenly spread along the Rasch developmental pathway. Revising the items in this manner should result in more precise measurement, which will lead to increases in the reliability and the separation coefficients.

Second, the number of items should be increased. Even if rewriting the current items in an effort to spread them evenly along the Rasch dimension were to be successful, there would still gaps in the range of the item difficulty estimates in which no items would be found. Thus, adding 10 to 15 more items would theoretically provide valuable information for the precise measurement of this sample. As with revision of the current items, the addition of items should also help to increase reliability and

Table 1
Item statistics for the Japanese metalinguistic knowledge test (measure order)

Entry Number	Total Score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	PT-Measure Corr.
3	3	100	4.43	.61	0.97	0.0	.24
6	31	99	1.15	.25	0.91	-0.7	.52
10	31	97	1.13	.25	1.16	1.4	.33
4	48	100	0.23	.23	0.93	-0.7	.51
7	48	97	0.16	.23	0.92	-0.8	.53
9	52	98	-0.01	.23	1.08	0.9	.42
5	66	97	-0.83	.25	0.99	-0.1	.47
1	80	100	-1.67	.28	0.84	-1.0	.52
2	84	100	-2.02	.31	1.08	0.5	.28
8	89	100	-2.57	.36	0.83	-0.6	.49
M	53.2	98.8	0.00	.30	0.97	-0.1	
SD	25.8	1.3	1.91	.11	0.10	0.8	

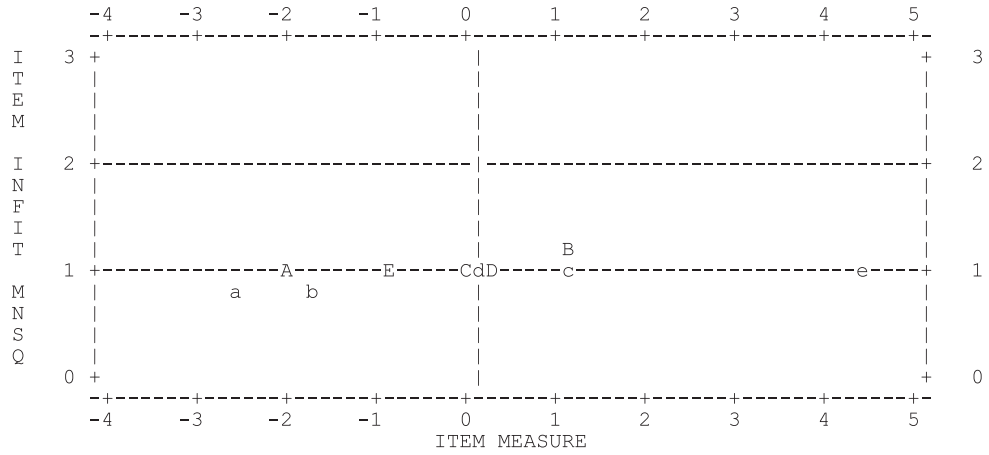


Figure 2. Plot of item infit mean-square values by item measure estimates

separation of the measures.

Finally, the fit statistics for the current set of items exhibit good fit with the Rasch model. These results suggest that the items are on construct and functioning to provide useful information in the measurement of the learners. Additional items, therefore, could be created in a similar style and format as the current set of items to create a focused, easy to administer set of items designed to test L1 metalinguistic knowledge.

Conclusion

The current study sought to investigate the functioning of a ten-item metalinguistic knowledge test that targeted Japanese metalinguistic terminology. The results of a Rasch analysis pointed to areas of the test that could be improved—most notably the gaps found along the Rasch dimension and the similar item difficulty estimates found for a few of the items. Implications drawn from these results focused on the need to rewrite some of the items, or aspects of the items, and on the need to include more items on the test to produce an instrument that is capable of reliably measuring and separating participants along the targeted latent variable.

References

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research, 1*, 93-121.

Elder, C., Warren, J., Hajek, J., Manwaring, D., & Davies, A. (1999). Metalinguistic knowledge: How important is it in studying a language at university? *Australian Review of Applied Linguistics, 22*, 81-95.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141-172.

Roehr, K. (2007). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics, 29*, 173-199.

Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.